Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most

promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the

customers with higher lead score have a higher conversion chance and the customers with lower lead

score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary

Step 1: Reading and Understanding Data

- The initial step involved loading the dataset and performing a thorough analysis to understand its structure, types of variables, and any inherent patterns. This included identifying the types of variables (categorical, numerical, etc.) and the initial inspection of missing values and potential data quality issues.

Step 2: Data Cleaning

- Handling Missing Values: Variables with a high percentage of missing (NULL) values were identified and removed from the dataset to maintain data integrity.

- Imputation: Missing values in numerical variables were imputed with the median values to minimize the effect of outliers and maintain the central tendency.

- Categorical Variables: New classification variables were created to handle missing values in categorical variables appropriately.

- Outlier Detection and Removal: Outliers were identified through statistical methods and visualizations, such as box plots, and removed to prevent skewing the analysis.

Step 3: Data Analysis

- Exploratory Data Analysis (EDA): Comprehensive EDA was conducted to understand the distribution, relationships, and patterns within the data. This included generating summary statistics, visualizing data distributions, and examining correlations between variables.

- Variable Reduction: Three variables that had only a single value across all observations were identified as non-informative and removed from the dataset.

Step 4: Creating Dummy Variables

- Encoding Categorical Variables: Categorical variables were transformed into dummy/indicator variables to make them suitable for machine learning models. This process involved creating binary columns for each category in the categorical variables.

Step 5: Train-Test Split

- Data Splitting: The dataset was divided into training and testing sets using a 70-30% split. This ensures that the model can be trained on a substantial portion of the data while reserving a separate portion for validation and testing.

Step 6: Feature Rescaling

- Min-Max Scaling: Numerical variables were rescaled using Min-Max Scaling to standardize the range of the data between 0 and 1. This step is crucial for algorithms that are sensitive to the scale of data.

- Initial Model Creation: An initial model was created using the stats model to provide a comprehensive statistical view of all parameters. This helped in understanding the significance and impact of each feature.

Step 7: Feature Selection Using RFE

- Recursive Feature Elimination (RFE): RFE was employed to iteratively select the top 20 most important features. This method helps in identifying the features that contribute the most to the predictive power of the model.

- Significance Testing: Using the statistical output (P-values), features were evaluated for their significance. Insignificant features were systematically dropped to refine the model, retaining only those that significantly contribute to the model's performance.

This detailed approach ensures a robust and well-prepared dataset for building accurate and reliable predictive models.